COURSE WORK 1: DATA MINING

MSc Data Science: Coventry University UK (**2024.2 BATCH**)

Name: P D Lolitha Lakshan Weerasinghe

Index Number: COMScDS242P-003

Question 1: Online Retail Data

GitHub Link - Code: <https://github.com/lolitha-lakshan-1/DataMining_assignment_1/blob/main/Assignment.Rmd>

## Task 1 – Data Understanding

As the initial task excel dataset was loaded to R Studio for preliminary analysis. In this step dataset was checked for,

1. How the dataset composed, what are the columns which exists?
   1. It was found all the said columns existed as mentioned which are,

InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, Country

1. Are the column data types being correct?
   1. It was noted that most column types contained correct data types including the “InvoiceDate”. However, it was noted “Stock Code” contained alpha numeric values rather than a 5-digit integral number.
2. What are the missing values?
   1. It was found that “Description” value is missing from 1454 transactions and “CustomerID” is missing from 135080 transactions.
3. What are the unique values in columns such as “Country” and “Stock Code”
   1. “Country” column contained all the valid country values except there were some instances of value being “Unspecified”. There are 446 transactions with country specified as “Unspecified.”
   2. “Stock Code” contained non product codes such as 'BANK CHARGES', 'POST', 'DOT', 'M', 'PADS', 'C2'
4. Are there any outliers in the numeric columns?
   1. “Quantity” column contained values ranging from -80995.00 to 80995.00. However, the mean is 9.55 and median being 3.00. There are 10624 transactions with negative quantities.
   2. “Unit Price” ranges from -11062.06 to 38970.00 while mean being 4.61 and median being 2.08. There are 2 transactions with negative unit prices.

Box Plot of Quantity prior to removing outliers.

A black and white diagram of a number

AI-generated content may be incorrect.

Box Plot of Unit Price prior to removing outliers.

A graph of a number of numbers

AI-generated content may be incorrect.

1. Are there any correlated columns?

A green squares with white text

AI-generated content may be incorrect.

According to the heatmap there were not feature correlations found.

### Data Cleansing

1. Removed transactions which contained negative value for Unit Price or Quantity fields. About 10626 transactions were removed.
2. Removed cancelled transactions which contained “C” in the InvoiceNo.
3. Removed transactions with non-product Stock Codes such as 'BANK CHARGES', 'POST', 'DOT', 'M', 'PADS', 'C2'. About 2324 transactions were removed from the dataset.
4. Removed transactions with Country specified as “Unspecified”. About 446 transactions were removed.
5. It was decided not to remove customer null transactions as it does not affect the basket values.
6. After cleansing there are 528513 transactions remaining on the dataset.

## Task 2 – Apriori Implementation

After the data cleansing dataset was grouped using the combination of Country and the InvoiceNo. This would result unique combinations of Country and InvoiceNo, removing any duplicates. Then it will count the number of distinct InvoiceNo per country, and sort them in descending order of count and would output the top three countries with the number of transactions.

top\_countries <- data %>%

distinct(Country, InvoiceNo) %>%

count(Country, sort = TRUE) %>%

slice\_head(n = 3) %>%

pull(Country)

Countries with most transaction counts:

1. United Kingdom 23494
2. Germany 603
3. France 461

Thereafter transaction list was creating using the StockCode and InvoiceNo from the top three countries.

Following code filters the dataset for a top country and keeps only unique combinations of InvoiceNo and StockCode to avoid duplicate items in the same transaction.

Thereafter it groups the items by invoice number to form a Apriori transaction list, which is converted into a transactions object required for Apriori analysis.

The Apriori algorithm is run with minimum support of 1%, confidence of 50%, and at least 2 items per rule. The resulting rules are sorted by lift to highlight the strongest associations.

for (country in top\_countries) {

cat("\n=====================\n")

cat("Running Apriori for:", country, "\n")

cat("=====================\n")

# Filter for country

country\_data <- data %>%

filter(Country == country) %>%

distinct(InvoiceNo, StockCode)

# Convert to transaction list and transactions object

trans\_list <- split(country\_data$StockCode, country\_data$InvoiceNo)

transactions <- as(trans\_list, "transactions")

# Run Apriori

rules <- apriori(transactions, parameter = list(supp = 0.01, conf = 0.5, minlen = 2))

rules <- sort(rules, by = "lift", decreasing = TRUE)

# Print readable rules

if (length(rules) == 0) {

cat("No rules found for", country, "\n")

} else {

# Convert rules to data frame

rule\_df <- as(rules, "data.frame")

rule\_df$LHS\_Desc <- get\_descriptions(lhs(rules), stock\_map)

rule\_df$RHS\_Desc <- get\_descriptions(rhs(rules), stock\_map)

cat("\nTop 10 rules by lift for", country, ":\n")

print(rule\_df %>%

select(LHS\_Desc, RHS\_Desc, support, confidence, lift) %>%

arrange(desc(lift)) %>%

head(20))

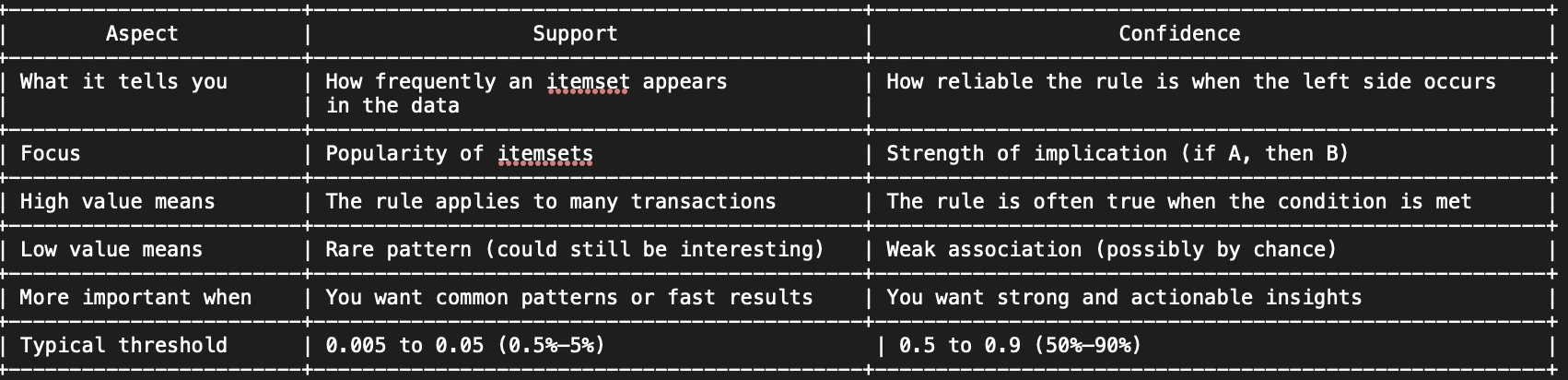
}

}

## Task 3 - Basket Analysis

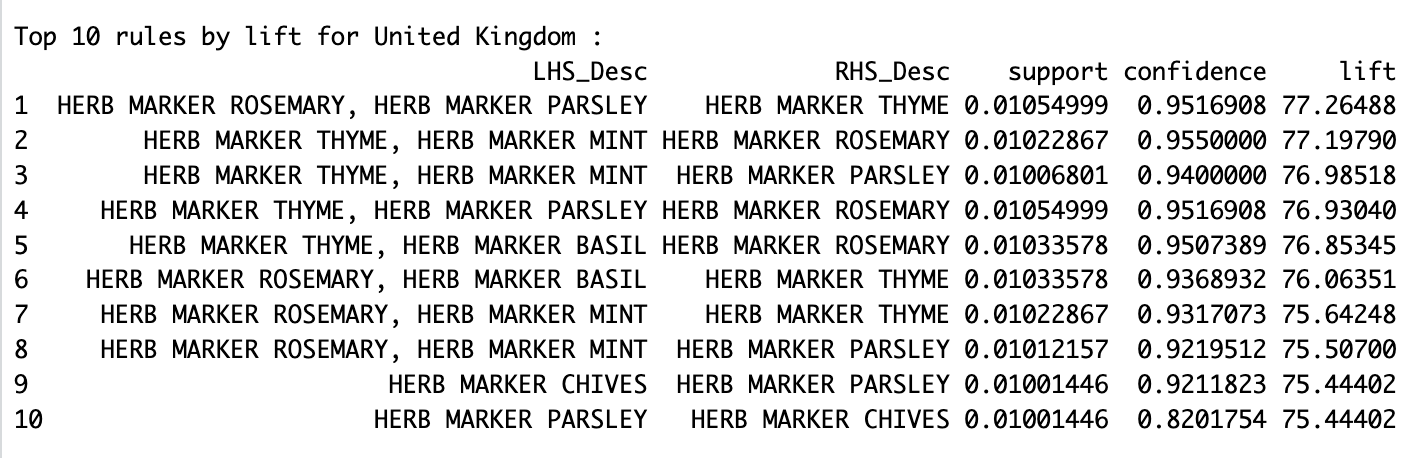
The initial Apriori run was done with a minimum support of 1%, confidence of 50%, and at least 2 items per rule. Although 70% - 80% will result strong reliable rules it will but it could lead to miss useful rules.

Support vs Confidence



Market Basket Analysis for United Kingdom

Apriori top 10 rules for parameters supp = 0.01, conf = 0.5, minlen = 2



Rule interpretations

1. It seems Customers who buy both ROSEMARY and PARSLEY almost always buy THYME too. Ideal for bundling.

2. If someone buys THYME and MINT, they are very likely also buy ROSEMARY. These three may form a strong trio of complementary items.

3. Similar to Rule 2 MINT and THYME pair strongly with PARSLEY. A different trio than Rule 2, but equally reliable.

4. Shows ROSEMARY is often bought when people buy PARSLEY and THYME. Reinforces prior rule.

5. Even substituting MINT/PARSLEY with BASIL still results in strong ROSEMARY prediction. ROSEMARY is a key crossover item.

6. ROSEMARY and BASIL jointly predict THYME. These three consistently show up across rules.

7. ROSEMARY + MINT again leads to THYME — confirms THYME’s strong linkage with most other herb markers.

8. MINT + ROSEMARY also predict PARSLEY with very high confidence.

9. CHIVES buyers very likely get PARSLEY — high lift suggests strong bond, despite single-item

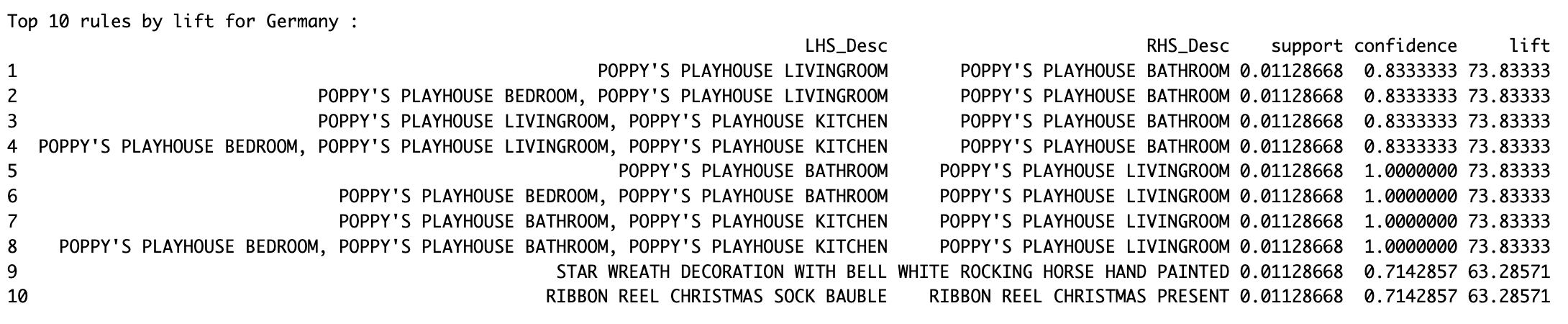
10. The reverse of Rule 9 — if someone buys PARSLEY, they frequently buy CHIVES too. Slightly lower confidence, same strong lift.

Overall, all these rules indicate extremely strong product associations within the “HERB MARKER” series of products. Almost all the rules’ projects confidence above 90% and lift above 75, which is exceptionally strong and incredibly high association strength. Seems like Herb markers are frequently purchased in combinations.

ROSEMARY, THYME, PARSLEY, and MINT consistently appear together, forming highly reliable combinations. Rules show that buying any two herbs almost guarantees purchase of a third. These findings are ideal for bundling and cross-selling.

Market Basket Analysis for Germany

Apriori top 10 rules for parameters supp = 0.01, conf = 0.5, minlen = 2 ordered for lift.



Overall, Germany shows strong associations within the "POPPY'S PLAYHOUSE" product line, especially between BATHROOM and LIVINGROOM sets.

Rules with 100% confidence suggest guaranteed purchases.

Holiday decorations like wreaths, baubles, and ribbon reels also exhibit strong bundling potential.

All 10 rules exhibit high lift values (63–74), indicating much higher co-occurrence.

Rule Interpretation

Rule 1 – 4

Customers who buy any playhouse room set almost always add the Bathroom set. Very strong cross-selling potential. With a confidence level of 83.33% and lift above 70% shows relatively strong and incredibly high association strength.

Rule 5–8:

With a confidence level of 100%, it seems Every customer who buys the BATHROOM set also buys the LIVINGROOM set. It's a guaranteed co-purchase highly valuable for promotions or recommendations.

Rule 9:

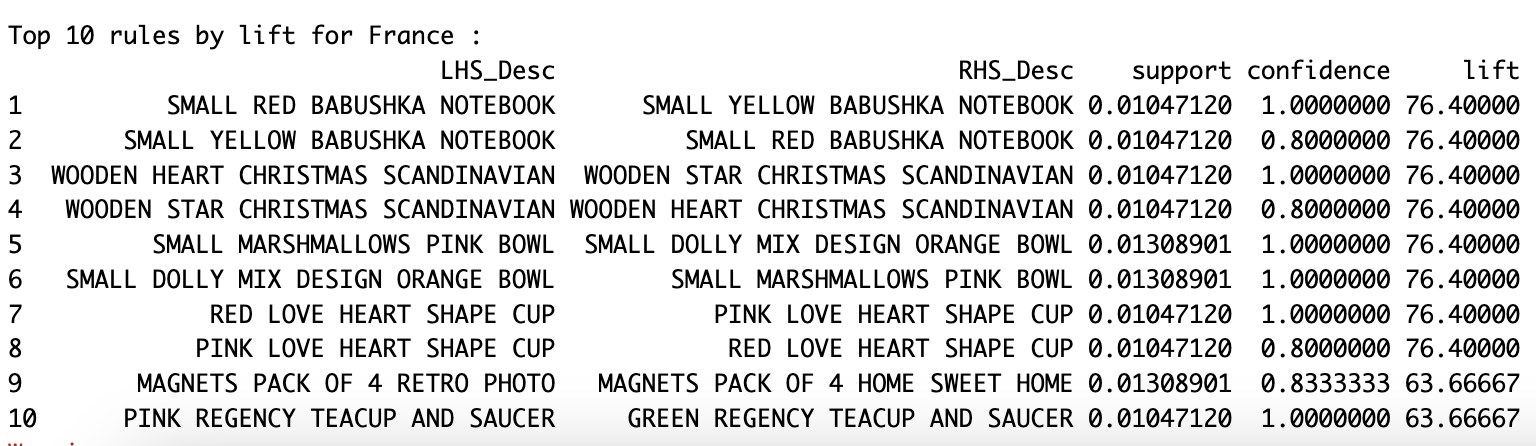
STAR WREATH DECORATION WITH BELL and WHITE ROCKING HORSE HAND PAINTED seems to be seasonal complementary items with strong association which are being bought together.

Rule 10

RIBBON REEL CHRISTMAS SOCK BAUBLE and RIBBON REEL CHRISTMAS PRESENT Christmas-themed ribbon items are highly associated. Strong opportunity for bundle deals during holiday seasons with a confidence level over 70%.

Market Basket Analysis for France

Apriori top 10 rules for parameters supp = 0.01, conf = 0.5, minlen = 2 ordered for lift.



Above rules depicts co-purchasing behavior among color or theme home variants. BABUSHKA notebooks, heart-shaped cups, and tea sets show 100% or near-perfect confidence, suggesting strong customer preference for matched or complementary products. These rules are ideal for bundling or gift-set curation.

Seasonal items like wooden Scandinavian decorations also pair reliably. Magnets and bowls reveal strong lift, reinforcing cross-sell value.

Rule Interpretation

Rule 1 – 2

SMALL RED BABUSHKA NOTEBOOK and SMALL YELLOW BABUSHKA NOTEBOOK seems to be highly correlated with a confidence value of 100% and lift of over 75%, therefore customers buying one BABUSHKA notebook very likely buy the other.

Rules 3 – 4

WOODEN HEART CHRISTMAS and WOODEN STAR CHRISTMAS highly correlated purchases with high confidence and lift values indicating frequent buys probably during the holidays. These two items can be seasonal and ideal to bundle together.

Rules 5 – 6

SMALL MARSHMALLOWS PINK BOWL and SMALL DOLLY MIX ORANGE BOWL seems to always be bought together which is ideal for bundling.

Rules 7 – 8

RED LOVE HEART CUP and PINK LOVE HEART CUP also being bought together just like above items. Mostly like as a pair or a set.

Rule 9

MAGNETS PACK OF 4 RETRO PHOTO likely to by HOME SWEET HOME product item.

Rule 10

PINK REGENCY TEACUP AND SAUCER and GREEN REGENCY TEACUP AND SAUCER a guaranteed co-purchase with confidence level of 100%. Seems like customers buying the pink variant always buy the green one, suggesting strong interest in color-matched tea sets.

Comparison with unit prices.